UC Riverside
Issues in Applied Linguistics

Title

A Comparison of the Effects of Analytic and Holistic Rating Scale Types in the Context of Composition Tests

Permalink

https://escholarship.org/uc/item/4dw4z8rt

Journal

Issues in Applied Linguistics, 11(2)

ISSN

1050-4273

Author

Carr, Nathan T

Publication Date

2000-12-30

DOI

10.5070/L4112005035

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

A Comparison of the Effects of Analytic and Holistic Rating Scale Types in the Context of Composition Tests

Nathan T. Carr
University of California, Los Angeles

This study examines how different composition rating scale types—analytic and holistic—can differentially affect the aspects of academic English ability measured in an ESL proficiency test battery. Specifically, the study addresses the following questions: (1) To what extent do holistic and analytic scales contribute differentially to total scores on a test of academic English ability? (2) To what extent does the test as a whole measure different aspects of language ability, depending on whether analytic or holistic composition scores are used? (3) To what extent does a particular rating scale type provide potentially useful information for placement or diagnosis, either alone or as part of a multi-component assessment? Multiple regression and exploratory factor analyses indicate that changing the composition rating scale type not only changes the interpretation of that section of a test, but may also result in total test scores which are no longer comparable.

INTRODUCTION

Rating scales, sometimes referred to as rubrics, can be powerful tools for systematically quantifying the performance of test takers. In the context of assessing the academic writing ability of nonnative speakers of English, they provide the potential for reliable scoring in a valid manner, rather than simply according to the personal idiosyncrasies of the rater.

Naturally, however, the use of rating scales involves certain tradeoffs as well. For example, Delandshere and Petrosky (1998) point out that there are inherent limitations involved in reducing complex performances to one or more numerical ratings and claim that such ratings therefore do a poor job of representing the constructs to which they are intended to correspond. Furthermore, they argue that the more complex a task is, the less easily it can be generalized to contexts outside that of the assessment. While there is probably some truth to these claims, we should not rush to throw out the baby with the bath. The testing context of interest here is the so-called one- or two-shot essay exam in which test takers write an essay in a single sitting (one "shot") and may also be given an opportunity to edit their writing (two "shots"). This testing context can reasonably be held as corresponding to the target language usage domain of academic writing in several respects. For example, many academic content courses require in-class essay tests, and most or all academic writing requires an introduction, a conclusion, and the use of rational arguments rather than emotional ones. In addition, in spite of any limitations inherent in reducing complex constructs to one or a few numbers, such

Issues in Applied Linguistics
ISSN 1050-4273
Vol. 11 No. 2, 207-241
© 2000, Regents of the University of California

ratings are too useful in describing writing not to be used, although perhaps we might do well to maintain a vigilant attitude towards potential problems with reliability and validity.

In considering whether to revise or replace a particular rating scale, there are two principal issues to address. The first is the identification of the target language usage domain to which test developers wish to generalize test results. The second issue is the question of what qualities of test usefulness (Bachman & Palmer, 1996) test developers wish to emphasize and how those qualities are and would be instantiated in the current and revised versions of the test. One quality of usefulness of particular interest in such situations is the contribution of a rating scale to the overall assessment process—in the present context, how well it describes test takers' performance, how it contributes to the test's predictive utility, how useful it is for diagnostic purposes, and what data it may potentially contribute for research.

These types of contributions define the research questions to be addressed by the present study, which examines how and to what extent changes in the composition rating scale used for the University of California, Los Angeles (UCLA) English as a Second Language Placement Examination (ESLPE) may have resulted in changes both in the constructs measured by the test and in the relative degrees of importance of its parts. Specifically, this study addresses the following questions:

(1) To what extent do holistic and analytic scales contribute to total scores on a test of academic English ability?

(2) Are reading, listening, and composition scores clearly distinct for both versions of the ESLPE, and if so, to what extent are they different? That is, to what extent does the test measure distinct aspects of language ability?

(3) To what extent does a particular rating scale type, or its subscales, provide potentially useful or distinctive information for diagnosis or research, either alone or as part of a multi-component assessment?

The answers to these questions are intended to contribute empirical evidence to the ongoing debate on the relative merits of analytic and holistic rating scales and be of practical local benefit in the ESLPE planning and decision-making process.

Conceptual Framework for Rating Scale Comparisons

One useful way of looking at types of rating scales (L. F. Bachman, personal communication, May 19, 1998) is to view them in terms of two dimensions. The first is that of *holistic* (sometimes called *global* or *unitary*) versus *analytic* (sometimes called componential) rating scales, which appears to be the principal philosophical division within the writing assessment community. Most, if not all, rubric validation research seems to relate to this issue as well. The second dimension is that of *generic* versus *tailor-made* rating scales, which refers to the division between highly task-dependent scales, such as primary and multiple trait, and more general ones, which are themselves usually held valid for only a limited range of contexts and task types. The dimension of generic versus tailor-made scales is

probably more a matter of degree than a true dichotomy, however, and although it clearly merits further investigation, it lies beyond the scope of this study.

Comparison of Holistic and Analytic Rating Scales

Holistic and analytic rating scales differ in that a holistic rating scale uses a single global numerical rating to rate a composition, while an analytic rating scale uses several subscales, which may or may not be summed or averaged together to form a composite total, to rate characteristics of a composition separately. Another distinction that has been made is that between holistic and enumerative approaches, with any nonenumerative scoring method labeled as holistic (e.g., Cooper, 1977; Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey, 1981). This sense of holistic, however, seems to have disappeared from use. A slightly different distinction is made by Hamp-Lyons (e.g., 1991), who distinguishes among holistic, primary trait, and multiple trait rating scales. Nevertheless, the most common current usage appears to be that employed in this paper.

Support for holistic scoring

White (1984), a principal proponent of holistic scoring, claims that while holistic scoring is not perfect, it offers a number of advantages over analytic scoring, the first being that "it has made the direct testing of writing practical" (p. 408). This argument is probably his strongest, as it is well known that simpler rating scales generally take less time to score by and therefore generally cost less to use than do more complicated ones. Furthermore, such scores tend to simplify the rater training process, which can be of considerable benefit in situations where the rater pool sees frequent turnover or when there is a frequent amount of rater retraining.[1]

The second advantage White claims for holisticism, however—that a single global rating tends to be more reliable than one from a rating scale consisting of several subscales—is somewhat more problematic. Reliable holistic rating scales can certainly be constructed, but all things being equal, rating scales with a greater number of subscales are generally seen as leading to greater overall consistency of scoring (see, e.g., Brown & Bailey, 1984; Hamp-Lyons, 1991). This perspective treats each subscale as being equivalent to a test item and is based upon the axiom that although merely adding items to a test will not necessarily add to its reliability, tests with more items or tasks are generally more reliable (Allen & Yen, 1979; Linn & Gronlund, 1995). In spite of this theoretical advantage for analytical scoring, however, Hamp-Lyons and Kroll (1997) concede that holistic ratings can achieve reliability rates close to those achieved by componential scales. One possible explanation for this is that it is difficult to create a workable rating scale using more than a few distinctive subscales.

In support of the construct validity of holistic scoring, Huot (1990) reports that the results of research on the process of rating using holistic rating scales indicate that "raters are most influenced by the content and organization of a

student's writing" (p. 207). This serves as evidence that holistic scoring, in the monolingual context at least, seems to be based upon some definable construct. Similarly, Tyndall and Kenyon (1996) report that a multi-faceted Rasch analysis indicates that a new holistic scale used in EFL placement at George Washington University appears to measure a single construct of writing ability. This single construct is implied to be something along the lines of "EFL Program course level," however, which raises a problem: While the test may satisfy its objective of placing students in the program, such scores cannot be interpreted as descriptions of writing ability but only as predictions of success or failure at a given level of study (Bachman, 1990).

Problems with holistic scoring

Huot (1990) considers the degree to which holistic scoring has been validated and finds that the emphasis in holistic scoring has strongly tended to be on reliability rather than validity, with the two sometimes even being equated. This, he claims, has led to an unsubstantiated general assumption of the validity of holistic scoring procedures. Most of the problems with holistic scoring center around its validity, with the most problematic issues perhaps being what it is that holistic scoring measures and whether holistic scores are able to adequately capture the whole of a written product in a single global rating.

With regard to the first concern, Hamp-Lyons (1990) argues that raters using holistic rating scales cannot agree on which essays are better than others, nor on what specifically makes one superior to another. In support of this argument, she points out that an interrater reliability coefficient of .70 means that raters are only in 49% agreement. Taken even further, the .90 correlation she reports for raters of the Michigan Test of English Language Proficiency, or MELAB (University of Michigan, n.d.; cited in Hamp-Lyons, 1990) indicates a 19% disagreement rate.

In addition, studies indicating that holistic scores represent a single construct, which requires agreement among raters as to what is being rated, turn out to be of questionable generalizability to the evaluation of nonnative speakers' academic writing. For example, while Huot's (1990) findings that raters using holistic rating scales focus on content and organization could be seen as lending support to the validity of holistic ratings, the generalizability of these results to the ESL context may be limited because the studies to which he refers do not appear to have differentiated between native and nonnative speakers of English. Similarly, a study by Vacc (1989) identified quality and development of ideas to be a significant predictor of holistic scores across all raters, but she herself cautions that her results may not be generalizable from low-ability monolingual eighth grade boys to other educational contexts. Even within that context, however, beyond their apparent agreement regarding one factor underlying writing ability, she finds that there tends to be little agreement among raters on what else is subsumed by holistic scores of writing ability.

In contrast to studies indicating the importance of content and discourse features in holistic scoring, Homburg (1984) found evidence supporting the idea that holistic rating scales discriminate between intermediate ESL students on the basis of linguistic accuracy. He does note, however, that the measures used in the study would probably be less important at high ability levels and might be unusable at lower levels. Through the use of a combination of five objective scoring measures (error rates, dependent clauses, words per sentence, coordinating conjunctions, and T-units), he was able to account for 84% of the variance in scores for these three levels. This finding indicates that trained raters using the holistic rating scale for the Michigan Test of English Language Proficiency (University of Michigan, n.d.; cited in Homburg, 1984) battery, a predecessor of the MELAB, probably focused primarily on language at certain ability levels. If linguistic accuracy and sophistication is to be the definition of academic writing ability, then this could be seen as a validation of the holistic approach. If the construct is defined as including something more, however, these results pose something of a problem for the validity of holistic scoring.

Moving to the second chief concern about holistic scoring—whether holistic scores are able to adequately capture the whole of a written product—it appears that holistic ratings of academic writing might suffice for native speakers, who have presumably attained a mastery of the linguistic forms of English and are more likely to have difficulty with other aspects of writing, such as content and organization. Such may not be the case for nonnative speakers, however, particularly those whose English ability is still a work in progress. Hamp-Lyons (1991, 1995) points out the problem that while some students may have comparably developed levels of grammar and organization in their writing, others may exhibit differing levels of performance and therefore cannot be accurately described by a single global score.

One example of this problem can be seen in Vaughan's (1991) study, in which negative comments expressed by at least three raters showed widely varying patterns. Disregarding comments on handwriting, disagreement with content, and offensiveness of content, and treating unclear content as being linguistically based, language-related comments accounted for as little as 33% (Student C) and as much as 100% (Student D) of all negative comments. Kroll (1990) provides a second example, noting that in a study of 100 essays by advanced academic ESL students, holistic ratings of discourse features (organization and coherence) were uncorrelated (ρ_{sb} = .083) with measures of syntactic accuracy. Furthermore, as might be expected, each band on the holistic rating scale showed wide variation in syntactic accuracy.

Another aspect of this concern over the descriptive adequacy of holistic ratings is raised by White (1984), who admits that a single score does little to provide a profile of a student's writing ability. Hamp-Lyons (1991, pp. 244-245; 1995, p. 761) and Hamp-Lyons and Kroll (1997, p. 29) take a similar position, adding that with holistic scoring, "diagnostic feedback is out of the question."

Going further, Hamp-Lyons (1991) also addresses White's (1984) claim that analytic scoring is reductionist, claiming that instead it is holistic scoring that is reductionist because it attempts to reduce "cognitively and linguistically complex responses to a single score" (p. 244). An idea of the problematicity of such reduction can be gained from Johns (1986), who in her discussion of coherence provides a window onto the simplification of the construct inherent in assigning a single global rating to writing. She describes coherence as an amalgam of features including cohesion, unity, register, a thesis, and logically related assertions. All of these factors are subsumed under a single term, which generally constitutes a single subscale at most and often only a component of one or two subscales, yet they by no means characterize the whole set of discourse features of writing ability.

Additional criticism of the descriptive adequacy of holistic rating comes in Hamp-Lyons' (1995) charge that it "fails as a qualitative research tool . . . [and] permits only quantitative research, limiting what can be known and permitting crude perceptions and categorizations" (p. 761). Finally, she adds that holistic rating can also prove inadequate when used for placing students into ESL programs offering a variety of course choices.

Support for analytic scoring

Some of the benefits of analytic scoring already alluded to above in the discussion of the weaknesses of holistic scoring include its potential to describe varying levels of ability across different features of a student's writing, such as grammar and organization, as well as to provide diagnostic feedback. While it may not be reasonable to expect much detailed diagnosis for students or teachers from an analytically scored composition test, the general comments provided in the various subscales' band descriptors could prove helpful to students.

Research by Fathman and Whalley (1990) into the effectiveness of teacher feedback on content (organization, description, coherence, and creativity) and grammatical errors indicates that general comments on content and specific comments on grammar can help students significantly improve their scores in both areas if they revise their essays. While students would not be likely to revise their essays from a university-wide ESL composition test, barring the addition of an editing task to the test, and would also not be likely to receive feedback much more detailed than the degree and overall types of their error patterns (e.g., cohesion, linguistic range and accuracy, organization), such feedback might nevertheless prove useful to them in their subsequent writing, whether for ESL or content area courses.

Setting aside for now questions of the usefulness of results, work by a number of researchers helps validate the notion of dividing writing ability into separate categories. For example, in a study by Cumming (1990), a multivariate analysis of the variance in ratings assigned by both experienced and novice raters showed significant main effects with no interactions, indicating the two groups "distinguished students' ESL proficiency and writing expertise as separate factors in their rating of the compositions" (p. 35). Cumming and Mellow (1996) provide rein-

forcement for this notion in a subsequent study, which also found no relationship between the two abilities.

Similarly, work by Weigle and Lynch (1996) on the validity of the then-recently revised ESLPE provided at least partial support for the validity of the constructs embodied in the test's three-subscale composition scoring system. Additional research by Milanovic, Saville, and Shuhong (1996) found that raters marking Cambridge First Certificate in English and Certificate of Proficiency in English (University of Cambridge Local Examinations Syndicate, n.d.; cited in Milanovic et al., 1996) compositions focused on different aspects of writing, specifically on "communicative effectiveness and task realization" versus vocabulary and content (p. 106), which further supports the notion that more than a single global rating is needed to adequately describe test taker performance.

A final factor militating towards the use of analytic scoring can be inferred from Weigle's (1994) study of the effects of rater training on the use of a three-component analytic scale. She observed that ratings showed differing sensitivities to training across subscales, an apparent indication that separate constructs were in fact being considered in the rating process.

Problems with analytic scoring

White (1984) raises what is probably the single most problematic issue for this method: Although analytic scoring should, in theory, allow for greater diagnostic information, he claims that there is no serious agreement as to "what, if any, separable sub-skills exist in writing" (p. 407). White's arguments are primarily oriented towards the monolingual context, however, and so may not be generalizable to the second or foreign language writing context. Furthermore, since the time of White's writing, work by Cumming (1990) discussed above indicates that raters in the ESL context are able to distinguish between language and writing abilities, which would indicate that at least a two-subscale analytic instrument is defensible. Still, the issue is far from settled. Is two the maximum number of distinguishable components of academic writing proficiency? Is there one best number for all contexts?

As an example of the difficulty in distinguishing where to draw the lines between components of writing ability, Huot (1990) claims that content and organization are the two most important factors in determining scores given by holistic raters. These two factors are sometimes difficult to distinguish, however, as was the case with the old version of the UCLA ESLPE composition rating scale. According to Weigle (S. C. Weigle, personal communication, June 5, 1998), this was an important factor in the decision to eliminate the former ESLPE analytic scale at UCLA. She also notes, however, that this may have been at least partially attributable to the wording of the specific rating scale. One might speculate that improved wording of the two subscales could have cured this problem, but it is also quite possible that content and rhetorical control may be intrinsically difficult to evaluate separately. Whichever the case, Hamp-Lyons and Henning (1991) also report

problems associated with distinguishing between multiple subscales and suggest reducing the number from the five or seven in their study to three, in part because of the heavy cognitive load which larger numbers of subscales impose on raters.

A second weakness associated with analytic scoring is that while more detailed, constraining rating scales increase overall interrater reliability, they take longer to use and therefore cost more (Popham, 1997).

One final possible weakness of analytic scoring is pointed out by Delandshere and Petrosky (1998), who opine that describing complex behaviors in terms of "a set of scores and generic feedback . . . falls short of providing useful representations and analyses" (p.16). They add that rating scales "are too generic for describing, analyzing, and explaining individual performances" (p. 21) and maintain that "substantive statements . . . [not] numerical ratings and generic feedback" (p. 16) are required.

BACKGROUND TO THE PRESENT STUDY: DEVELOPMENT OF THE ESLPE COMPOSITION RATING SCALES

Prior to 1989, the UCLA ESLPE composition test was used only for research purposes and as a tiebreaker when students were at the cut point between levels. In the summer of 1990, as a part of an overall test revision process, graduate students Charlene Polio and Sara Cushing Weigle were hired as research assistants to develop a new composition test for inclusion in the ESLPE. Working primarily with Christine Holten, the supervising lecturer for composition in the UCLA ESL Service Courses, they developed an analytic model with three subscales: content, rhetorical control, and language, with language double-weighted (C. Holten, personal communication, April 28, 1998). Brian Lynch, who at that time was both Director of the ESL Service Courses and Director of the ESLPE, adds that the reason for choosing this structure was "to have the scale parallel the aspects of writing that were emphasized in the ESL Service Course curriculum" and that the emphasis on the language subscale was "because the ESLPE was first and foremost a *language* test. Experience showed that if too much relative weight was given to content and organization, it would not reflect the range of students' language proficiency levels" (B. Lynch, personal communication, June 3, 1998).

By 1996, problems with the scale had become apparent: Specifically, raters had difficulty distinguishing content and organization, and the scale seemed cumbersome in operational use. An examination of actual practice revealed that when students or graduate teaching assistants in the ESL Service Courses complained about student placements, the main factor considered in the decision as to whether to re-place the student was his or her language score, not that for content or organization. The decision was therefore made to begin work on developing a new rating scale based entirely upon language. It was first operationally used during the 1996-97 academic year and remained unchanged at the time these analyses were performed in 1998 (C. Holten, personal communication, April 28, 1998).

The rating scale underwent minor adjustments for the first time in September 2000. All three rating scales—the analytic scale and both the original and revised language-only holistic scales—are included in the Appendix.

METHODOLOGY

The study used an ex post facto correlational design (Gay, 1992; Isaac & Michael, 1995) to investigate the effects of changing the ESLPE composition rating scale from an analytic to a holistic model. Also known as *retrospective* or *causal-comparative*, this design was appropriate because the change being studied had already taken place and because the analyses used investigated the relationships between variables based on correlation coefficients.

Data

The data used for this study came from the Fall 1995 administration of the ESLPE. The listening and reading portions of the test were scored objectively, and the compositions were scored using the analytic rating scale. These ratings are generally performed by lecturers and graduate teaching assistants from the ESL Service Courses the day after the test is administered. After all the essays from a given administration have been rated once, they are randomly redistributed for second ratings during the same rating session.

As a part of her dissertation research as a student in the UCLA Graduate School of Education and Information Science, Cynthia Taskessen obtained 94 of

Table 1 List and Description of Variables Used in the Study

Variable	Description
COMPH1	Holistic Composition score from first rating
COMPH2	Holistic Composition score from second rating
COMPA1	Total analytical composition score from first rating
COMPA2	Total analytical composition score from second rating
CONT1	Composition content rating from first rating
CONT2	Composition content rating from second rating
ORG1	Composition organization rating from first rating
ORG2	Composition organization rating from second rating
LANG1	Composition language rating from first rating
LANG2	Composition language rating from second rating
LIST	Total listening score
READ	Total reading score

the compositions from this administration of the test and had 83 of them rescored using the new holistic rating scale. These rescorings were performed by graduate teaching assistants from the ESL Service Courses. A list of the variables used in the study, along with a brief description of each, is provided in Table 1 on the preceding page.

Analyses

The statistical methodologies most appropriate to the data and the research questions were multiple linear regression and exploratory factor analysis. They were performed in that order, as it corresponded to that of the research questions.

Multiple linear regression

The first stage of the analysis of these data involved the construction of a series of multiple linear regression models using SPSS for Windows Release 8.0.0 (SPSS Inc., 1997) in which the overall ESLPE scores were regressed on the scores from the component parts. This was done in order to determine how changes in the composition rating scale might influence the relative contributions made by various ESLPE component scores, particularly the composition ratings, to the overall test scores. Although the terms "predict" and "predictor" are used in describing these regression analyses and their results, it should be made clear that no attempt to actually predict overall scores was intended. Such an approach would hardly be useful or appropriate, given that the subscores should account entirely for the overall score. Rather, these regression analyses were intended solely for the purpose of determining the relative importance of each component score in determining overall test scores.

Because the overall scores were computed differently using the different rating scales, it was necessary to use separate independent variables. Total scores, rather than placement levels, were used for two reasons. First, using placement levels would have resulted in dependent variables with only six values, which would be equivalent to using ANOVA, and would not have allowed capturing all of the variation attributable to the independent variables. Second, using placement levels would have confounded the effects of rating scale variation with cut scores.

Three regression models were constructed using forward stepwise regression and listwise deletion of cases with missing data, as follows:

> Model 1: The exam scored using holistically rated compositions was used to construct Regression Model 1: SCOREH = LIST + READ + COMPH (total score = listening score + reading score + holistic composition rating scale score).

> Model 2: Exam scores based on analytically rated compositions were used to construct Regression Model 2: SCOREA = LIST + READ + COMPA (total score = listening score + reading score + analytic composition rating scale score).

Model 3: Regression Model 3 considered the analytic rating scale by its component subscale scores of content, organization, and language: SCOREA = LIST + READ + CONT + ORG + LANG (total score = listening score + reading score + content subscore + organization subscore + language subscore).

Regression assumptions were confirmed for all three models using normal P-P plots of the regression standardized residuals; bivariate scatterplots of the standardized residuals and dependent variables, predicted and observed values of the dependent variables, and dependent and independent variables; and histograms of the residuals.

The relative importance of the predictors was examined by comparing their standardized regression weights and "R^2 change-if-last" values. The latter were obtained separately for each variable by entering all other variables in one block and the variable in question in a second block, with forced entry used in both blocks. These R^2 change-if-last values show the contribution of each variable to the overall model when all variables are included and can be interpreted as the proportion of the variance in the dependent variable uniquely attributable to the independent variable in question.

Exploratory factor analysis

The data were further analyzed by performing an exploratory factor analysis of the variables in each of the three regression models. All factor extractions were performed with SPSS for Windows Release 8.0.0 (SPSS Inc., 1997) using principal axis factoring, which uses squared multiple correlations on the diagonals of the correlation matrices as the initial estimates of the communalities. All composition ratings were analyzed using the scores given by individual raters in an attempt to reduce the risk of factor underdetermination. The SPSS default option of using Pearson *r* as the correlation coefficient was chosen, as reading and listening scores were interval data and the rating scale data had relatively normal distributions, with no variable having a skewness with absolute value greater than .377 (CONT1) or kurtosis with absolute value greater than .699 (COMPH2). The combined correlation matrix for all three models is given in the Appendix in Table A3. Principal factor analysis was chosen over principal components because the latter does not permit accurate estimates of the number of common factors (Carrol, 1993; Comrey & Lee, 1992). Following examination of the scree plot for a preliminary estimate of the correct number of factors to extract, a methodology similar to that described by Comrey & Lee was used. All common factors with positive eigenvalues were initially extracted, those unrotated factors with no loadings of at least .20 were discarded, and the extraction process was rerun, with the number of factors to be extracted set equal to the number of factors retained from the previous extraction.

The resulting matrix was rotated using the Equamax algorithm with Kaiser Normalization to approximate Comrey's Tandem Criteria rotation method (Comrey & Lee, 1996). All factors with no loadings of at least .30 or higher were discarded, and the entire extraction and rotation process repeated until the number of factors

to be retained had stabilized. The extraction was then performed one additional time, with the resulting factor matrix rotated obliquely using Promax with Kaiser Normalization. Solutions were then inspected to ensure that blind rotation to simple structure had not yielded uninterpretable complex-composite factors. To confirm that the correct number of factors had indeed been extracted for a given model, the scree plot was reexamined; the factor structure was considered to ensure it was interpretable; and full solutions were attempted, and rejected, with one more and one fewer factors extracted.

The extraction and rotation solutions were iterated to convergence, with the maximum number of iterations set at 200. Missing data were deleted listwise, except in the case of Model 3, in which iterations sufficient to allow convergence generated one or more variables with a communality greater than 1. Missing values were thus replaced with the corresponding mean scores. This problem raises the potential weakness of the study, that is, the size of the sample. Although 100 subjects is normally considered the minimum for obtaining reliable factor analytic results, the study was performed on an intact dataset, which prevented the reconstruction of missing data. The use of such a dataset was necessitated by the difficulty of obtaining scores for test takers rated using two rating scales on the same test. Further research exploring the factor structures of the two versions of the ESLPE with larger samples, which should better ensure factor stability, is of course desirable.

RESULTS

The results of the regression and factor analyses indicated an interesting combination of similarities and differences between test scores calculated using the two composition rating scales.

Multiple Linear Regression

Both standardized regression weights and R^2 change-if-last values indicated identical orderings of the independent variables for all three models, with the exception of content and organization scores in Model 3, for which the standardized regression coefficients differed by .005, but the R^2 change-if-last values were identical. The order in which predictors entered the model matched the rank orders of their standardized regression coefficients in every case. The correlation matrix and descriptive statistics for Model 1 are given in the Appendix in Table A1, and for Models 2 and 3 combined in the Appendix in Table A2.

Model 1: Holistic scale with listening and reading

The highest correlation between two predictors (r = .488 for listening and reading—see Appendix, Table A1) was not large enough to raise concerns over multicollinearity, and the minimum tolerance turned out to be .712 (see Table 2). All three predictors were found to be significant, and examination of the standardized regression coefficients and R^2 change-if-last values indicated that the most

important predictor of total ESL Placement Exam score as it was calculated using the holistic composition rating scale (SCOREH) was the composition subtest score (COMPH), followed in order of importance by scores for reading (READ) and listening (LIST). All values for *t* should equal infinity in this model, as they are equal to the raw regression coefficient divided by its standard error (Lewis-Beck, 1980; Neter, Kutner, Nachsheim, & Wasserman, 1996), which is zero in this case. The likely explanation for their large but finite magnitudes (see Table 2) is rounding error.

Table 2: Regressions of Placement Score on Component Subscores, Model 1

	Predictors		
	LIST	READ	COMPH
b	1.000	1.000	1.000
Std. Error	.000	.000	.000
ß	.347	.385	.569
R^2 Change if Last	.092	.106	.271
t	343,628,167	367,992,375	589,399,115
$p \leq$	.000	.000	.000
Tolerance	.768	.712	.837

$R^2 = 1.000$, $R^2_{adj} = 1.000$, $F = 2.70 \times 10^{17}$, $p \leq .000$

Models 2 and 3: Analytic scale with listening and reading

The correlation matrix for these models is given in the Appendix in Table A2. None of the correlations between the predictor variables for Model 2 were large enough to be of concern, but the correlation between content and organization ($r = .820$) was potentially problematic for Model 3. The minimum tolerance in Model 3 was .282 for organization (see Table 4), however, well above the .01 criterion for exclusion of the variable (Neter et al., 1996).

As can be seen in Table 3 on the following page, for Model 2 all three predictors were found to be significant. Examination of the standardized regression coefficients and R^2 change-if-last values indicates that for this model, the most important predictor of the total ESL Placement Exam score as it was calculated using the analytic composition rating scale (SCOREA) was the reading subtest score (READ), followed in order of importance by listening (LIST) and composition (COMPA) scores. As with Model 1, *t* values probably differ from infinity because of rounding error.

In Model 3, all predictor variables were again found to be significant, with their order of importance being reading, listening, language, organization, and content. As Table 4 indicates, however, although the standardized regression coef-

Table 3: Regressions of Placement Score on Test Subscores, Model 2

	Predictors		
	LIST	READ	COMPA
b	1.000	1.000	1.000
ß	.000	.000	.000
Std. Error	.442	.512	.374
R^2 Change if Last	.148	.189	.127
t	375,045,160	423,902,666	647,572,289
$p \leq$	.000	.000	.000
Tolerance	.759	.720	.906

$R^2 = 1.000$, $R^2_{adj} = 1.000$, $F = 3.17 \times 10^{17}$, $p \leq .000$

ficients for the last two variables differed slightly, their R^2 change-if-last values were identical to three decimal places, indicating little difference in the relative importance of the two variables. This is not surprising in light of the problem mentioned above, that is, that raters often had difficulty separating the two scores.

Table 4: Regressions of Placement Score on Test Subscores, Model 3

	Predictors				
	LIST	READ	CONT	ORG	LANG
b	.907	.985	.788	.783	2.155
Std. Error	.047	.043	.264	.263	.221
ß	.401	.505	.095	.100	.215
R^2 Change if Last	.115	.168	.003	.003	.030
t	19.128	23.111	2.982	2.982	9.732
$p \leq$	.000	.000	.004	.004	.000
Tolerance	.719	.661	.308	.282	.643

$R^2 = .972$, $R^2_{adj} = .971$, $F = 616.587$, $p \leq .000$

Exploratory Factor Analyses

Probably the most striking result of the regression analyses was the way in which altering the composition rating scale changed the order of importance of the various subscores in predicting total score. This, along with the seemingly low correlation of .497 between scores on the two scales, suggested the possibility that the two versions of the test represented by Models 1 and 2 might actually measure different constructs. To test this possibility, exploratory factor analyses of the independent variables in the three models seemed appropriate.

Model 1: Holistic scale with listening and reading

As with the multiple regression analyses above, Model 1 examined holistic composition ratings along with listening and reading scores. The communality estimates, summary of total variance explained by initial eigenvalues and extraction eigenvalues (sums of squared factor loadings), table of factor loadings, and correlations between oblique common factors are presented in Tables 5-8 and the in the scree plot in the Appendix in Figure A1. Two factors were extracted, with a good approximation of simple structure following oblique rotation. Factor I emerged as a holistic rating of written linguistic accuracy and Factor II as a receptive language ability rating. It would not appear reasonable to interpret Factor II as a selected response method factor, given its correlation (r = .481) with Factor I (see Table 8). The solution accounted for 65.543% of the variance in the model (see Table 6). The single-factor solution proved unsatisfactory because it yielded reduced loadings for all variables, indicating that more than one factor is necessary to describe the latent space. The three-factor solution was also rejected, as it produced a minor uninterpretable composite factor.

Table 5: Initial and Extraction Communalities, Model 1

Variables	Initial	Two Factors
SCORH1	.685	.839
SCORH2	.671	.801
LIST	.252	.458
READ	.298	.523

Table 6: Total Variance, Model 1

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
I	2.293	57.335	57.335	2.011	50.281	50.281
II	1.020	25.490	82.825	.610	15.262	65.543
III	.508	12.705	95.530			
IV	.179	4.470	100.000			

Table 7: Table of Factor Loadings, Model 1

Holistic Rating Scale		
	Factors	
Variables	I[a]	II[b]
SCORH1	**.900**	.032
SCORH2	**.908**	-.028
LIST	-.032	**.692**
READ	.039	**.704**

Note: Pattern Matrix
[a]Factor I is interpreted as a holistic rating of written linguistic accuracy.
[b]Factor II is interpreted as a receptive language ability rating.

Table 8: Correlations Between Oblique Common Factors, Model 1

Holistic Rating Scale		
	I	II
I	1.000	
II	.481	1.000

Models 2 and 3: Analytic scale with listening and reading

Model 2, which considered overall analytic composition ratings (content, organization, and doubled language scores for each rater) with reading and listening scores, yielded a two-factor solution similar to that for Model 1. This solution is presented in Tables 9-12, and its scree plot in the Appendix in Figure A2. This result was somewhat surprising, given the differences in relative importance of parts for the two models observed in the multiple regression analyses above. The interpretation of Factor I (composition) differs from that in Model 1, of course, because the two rating scales operationalize the construct of academic composition ability differently: One includes content and organization, while the other does not. The factors were correlated at .534 (see Table 12), and the solution accounted for 60.596% of the variance in the scores (see Table 10). As with Model 1, the single- and three-factor solutions proved unsatisfactory. In the single-factor solution, only one variable showed increased loading, while the others all decreased. A three-factor solution produced a minor uninterpretable composite factor.

Table 9: Initial and Extraction Communalities, Model 2

Variables	Initial	Two Factors
SCORA1	.526	.683
SCORA2	.478	.727
LIST	.277	.593
READ	.297	.421

Table 10: Total Variance, Model 2

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
I	2.273	56.837	56.837	1.899	47.475	47.475
II	.917	22.913	79.750	.525	13.121	60.596
III	.511	12.764	92.514			
IV	.299	7.486	100.000			

Table 11: Table of Factor Loadings, Model 2

Analytic Rating Scale		
	Factors	
Variables	I[a]	II[b]
SCORA1	**.750**	.130
SCORA2	**.895**	-.084
LIST	-.069	**.804**
READ	.111	**.583**

Note: Pattern Matrix
[a]Factor I is interpreted as a rating of academic composition ability, operationalized as including content, organization, and linguistic accuracy.
[b]Factor II is interpreted as a receptive language ability rating.

Table 12: Correlations Between Oblique Common Factors, Model 2

Analytic Rating Scale		
	I	II
I	1.000	
II	.534	1.000

Model 3, on the other hand, which included the subscores for the analytic rating scale, produced a five-factor solution, which accounted for 70.148% of the variance in the model. The results of this analysis are detailed in Tables 13-16, with the scree plot contained in the Appendix in Figure A3. Factors I and II (r = .674—see Table 16) represent content and organization scores for second and first raters, respectively (see Table 15). This appears to be uninterpretable at first glance, since first and second ratings were performed during the same rating session by raters drawn from a common rating pool. However, this most likely reflects the rating situation during that administration of the ESLPE—for example, it may be indicative of rater fatigue setting in during the second ratings. The fact that content and organization load on the same factor serves to highlight the problem mentioned above—that raters frequently had difficulty applying these two parts of the analytic rating scale.

Table 13: Initial and Extraction Communalities, Model 3

Variables	Initial	Five Factors
CONT1	.582	.782
CONT2	.604	.782
ORG1	.617	.813
ORG2	.637	.759
LANG1	.345	.462
LANG2	.424	.673
LIST	.315	.642
READ	.373	.698

Table 14: Total Variance, Model 3

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
I	3.777	47.213	47.213	3.503	43.788	43.788
II	1.299	16.242	63.455	.990	12.370	56.159
III	.839	10.483	73.938	.478	5.975	62.134
IV	.624	7.803	81.741	.374	4.671	66.805
V	.578	7.230	88.972	.267	3.343	70.148
VI	.420	5.251	94.223			
VII	.239	2.987	97.210			
VIII	.223	2.790	100.000			

Second ratings of ESLPE compositions are generally performed after all first ratings have been completed. This may have resulted in some sort of ordering effect, or raters may have taken a break, or not taken a break and grown fatigued. Reconstructing the rating process years later is extremely difficult, and this is one of the limitations often found in ex post facto research designs. Furthermore, the group of raters for this administration of the ESLPE contained a number of first-time graduate teaching assistants (Christine Holten, personal communication). Content and organization were often difficult for raters to separate in this rating

Table 15: Table of Factor Loadings. Model 3

	Factors				
Variables	I[a]	II[b]	III[c]	IV[d]	V[e]
CONT1	.054	**.874**	-.103	-.150	.215
CONT2	**.895**	-.016	-.036	.020	.024
ORG1	-.002	**.809**	.124	.157	-.193
ORG2	**.818**	.056	.052	.013	-.059
LANG1	-.045	-.002	**.665**	.097	-.010
LANG2	.063	-.009	**.797**	-.132	.097
LIST	.033	-.026	-.036	**.726**	.159
READ	-.027	.036	.077	.190	**.685**

Note: Pattern matrix.
[a]Factor I is interpreted as a combination of content and organization on the first essay rating.
[b]Factor II is interpreted as a combination of content and organization on the second essay rating.
[c]Factor III is interpreted as a rating of written linguistic accuracy.
[d]Factor IV is interpreted as a rating of listening ability.
[e]Factor V is interpreted as a rating of reading ability.

Table 16: Correlations Between Oblique Common Factors. Model 3

	I	II	III	IV	V
I	1.000				
II	.674	1.000			
III	.550	.577	1.000		
IV	.226	.390	.490	1.000	
V	.194	.312	.377	.414	1.000

scale, something which helped motivate the change to a holistic rating scale in 1996. As a result, it makes sense that ratings on these two subscales might have been particularly susceptible to systematic influence by extraneous factors, such as rater inexperience or rater fatigue, and that the scores for first and second raters on these two subscales in particular, as opposed to separate scores for each of the two subscales, might emerge as separate factors. It therefore seems reasonable to

interpret content and organization as a single construct until further research can confirm or disconfirm this hypothesis.

Factors III (language), IV (listening), and V (reading) are rather more straightforwardly interpretable. Although factors with only one loading are not generally desirable, the solution approaches simple structure well while remaining interpretable. On the other hand, three- and four-factor solutions each yielded a variable with a communality greater than one, which halted the extraction process. A six-factor solution produced a minor uninterpretable composite factor. Finally, examination of the scree plot, presented in the Appendix in Figure A3, further indicates the appropriateness of this solution. In summary, given the interpretability of the five-factor solution, examination of the eigenvalues and scree plot, and the problems encountered with alternative solutions, five factors best describe the latent space constituted by the ESLPE when subscores on the analytic rating scale are considered.

CONCLUSIONS

The results of the multiple regression and exploratory factor analyses yielded two main findings. First, changing the composition rating scale resulted in an alteration in the factor structure of the test sufficient to change the relative importance of its components, despite the fact that no other sections were changed. The nature of this change was such that the emphasis of the ESLPE changed from the receptive to the productive modality of language use. Second, because of this change, test scores derived using the two rating scales are not comparable, despite the fact that they are intended to measure the same construct, academic English language ability. This lack of comparability is not apparent on the surface, however, and is obscured by the .880 correlation between scores on the two versions of the tests.

The implications of these results are perhaps best framed in terms of the research questions posed above. Research Question 1 addressed the extent to which holistic and analytic scales contribute to total scores, and thereby to placement in an academic ESL program. The results of the present study show that changing rating scale types has the potential to fundamentally alter the overall emphasis of the test, even if other components are left untouched. In this case, changing the scale transformed the test from one focusing on reading, listening, and composition, in that order, to one focusing on composition, reading, and listening.

Research Question 2 dealt with the extent to which reading, listening, and composition scores are clearly distinct for the ESLPE, and therefore to what extent the test measures distinct aspects of language ability. The findings above indicate that listening and reading scores are not distinct from each other except when analytic composition scores are analyzed by their component subscores. This, along with the fact that the scores are correlated, suggests the possibility that a higher-order factor may underlie these primary trait factors. Further research, perhaps

employing structural equation modeling, would be necessary to investigate this question. One hypothesis which such a follow-up study might test would be that although differences exist between the two types of selected response sections, they emerge as loading on separate factors at a lower level in the factor hierarchy than do receptive language ability and composition ability (as measured with either rating scale).

Research Question 3 involved the extent to which a particular rating scale type or its subscales provide potentially useful or distinctive information for diagnosis or research, either alone or as part of a multi-component assessment. Findings here are somewhat more difficult to interpret than is the case for the first two research questions. Obviously, if composition scores are reported as single composite numbers, those derived from different rating scales will provide different types of information. The .497 correlation between scores on the two composition rating scales examined in this study illustrates this point, and should come as no surprise, as each rating scale is a different operationalization of the construct of academic writing ability. The difference is principally one of focus: Holistic scores provide an assessment of a single construct, whereas composite scores from an analytic rating scale conflate the information from several constructs. In the latter case, a single number can at best tell test users that a given test taker has all high or all low scores on the various components of the rating scale; a mid-level overall score, on the other hand, is ambiguous, as it could mean average ability levels across the board, or high scores in some areas combined with low scores in others.

The analyses detailed above demonstrate that if only overall composition scores are considered, neither of the rating scales analyzed in this study provides more information than the other, as their factor structures are essentially the same. In fact, when composition ratings are considered as unitary composite scores, the exploratory factor analyses above indicate that the two selected-response portions of the test do not provide separately interpretable scores. That is, both the listening and reading scores load primarily on the same factor. On the other hand, when the analytic scores are presented in terms of their subscores, much more information becomes available, with four interpretable constructs represented in the factor structure of the test. Therefore, if only a single composition score is to be used in research, placement, or diagnosis, it might be best to use a holistic rating scale in testing situations with factor structures similar to those of the two versions of the ESLPE. In contrast, when the various subscores of an analytic rating scale are to be used, such scores clearly provide a greater amount of distinctive information regarding test takers' abilities. One important deciding factor—perhaps the most important—should then be the degree to which the additional information is useful to test users.

In the context of the ESLPE or similar tests with similar factor structures, it is likely that for both research and diagnostic purposes, the analytic rating scale can clearly provide more potentially useful information. Provided that test scores are considered in terms of their component subscores, test users can be provided

information regarding four aspects of language ability, while test scores containing a single composition score only provide information about two. For use in ESL course placement alone, on the other hand, the holistic rating scale might be more useful in many situations, given that language ability level is "probably the most common criterion for grouping in such programs" (Bachman, 1990, p. 58) and that any differences between an individual test taker's subscores would probably be obscured by summing. This presupposes that subscores are weighted equally; if a composite test score were computed using weighted subscores, however, an analytic rating scale might prove more useful, provided that content and organization were intended to be viewed as a part of the construct of academic language ability.

Finally, a replication of these analyses using structural equation modeling might prove of interest. The application of this methodology would allow explicit testing of the degree of fit of the factor structures of the three models. It would also permit the investigation of additional questions, particularly the comparison of the models described above with alternative factor structures.

ACKNOWLEDGEMENTS

I wish to thank Cynthia Taskessen for graciously permitting the use of the dataset used in this study. I would also like to thank Dr. Lyle Bachman, of the University of California, Los Angeles, and two anonymous reviewers for their useful comments on previous drafts of this article. Of course, any remaining deficiencies are my own responsibility.

NOTES

[1] I am indebted to an anonymous reviewer for pointing this out.
[2] For ease of reference, I will henceforth use *organization* in place of the more cumbersome *rhetorical control*.

APPENDIX
TABLES A1-A3 &FIGURES A1-A3

Table A1: Correlation Matrix for the Holistic Rating Scale

	N	$\overline{X}$	SD	LIST	READ	COMPH	SCOREH
LIST	83	19.494	4.121	1.000			
READ	83	29.337	4.583	.488	1.000		
COMPH	83	22.807	6.772	.286	.386	1.000	
SCOREH	83	71.639	11.892	.690	.767	.817	1.000

Table A2: Correlation Matrix for the Analytic Rating Scale

	N	$\overline{X}$	SD	LIST	READ	COMPA	CONT	ORG	LANG	SCOREA
LIST	94	19.362	4.077	1.000						
READ	94	29.372	4.731	.488	1.000					
COMPA	94	23.457	3.457	.201	.300	1.000				
CONT	94	5.878	1.118	.228	.320	.702	1.000			
ORG	94	5.739	1.175	.284	.248	.743	.820	1.000		
LANG	94	5.899	.923	.324	.393	.788	.451	.528	1.000	
SCOREA	94	72.191	9.233	.767	.840	.617	.527	.531	.639	1.000

Table A3: Correlation Matrix for Factor Analyses

	N	X	SD	COMPH1	COMPH2	COMPA1	COMPA2	CONT1	CONT2	ORG1	ORG2	LANG1	LANG2	LIST	READ
COMPH1	85	3.788	1.235	1.000											
COMPH2	84	3.786	1.120	.829	1.000										
COMPA1	94	23.553	3.588	.502	.507	1.000									
COMPA2	94	23.277	4.060	.389	.423	.688	1.000								
CONT1	94	5.947	1.213	.248	.200	.753	.560	1.000							
CONT2	94	5.809	1.354	.211	.218	.520	.816	.518	1.000						
ORG1	94	5.755	1.284	.275	.323	.823	.619	.710	.523	1.000					
ORG2	94	5.723	1.363	.230	.273	.572	.830	.537	.763	.575	1.000				
LANG1	94	5.926	1.008	.541	.552	.802	.494	.287	.281	.401	.329	1.000			
LANG2	94	5.872	1.100	.444	.476	.596	.829	.382	.417	.465	.442	.535	1.000		
LIST	94	19.362	4.077	.293	.278	.372	.256	.219	.180	.333	.175	.318	.253	1.000	
READ	94	29.372	4.731	.406	.345	.396	.327	.367	.199	.263	.180	.317	.369	.488	1.000

FIGURE A1 SCREE PLOT (MODEL 1)

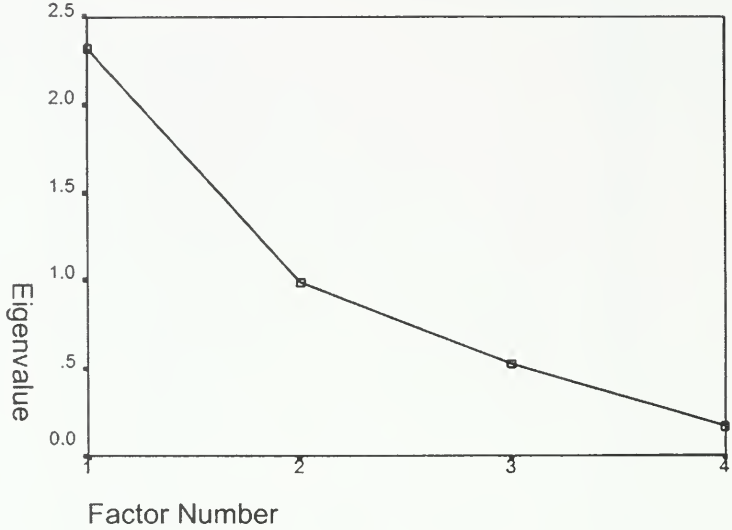


FIGURE A2 SCREE PLOT (MODEL 2)

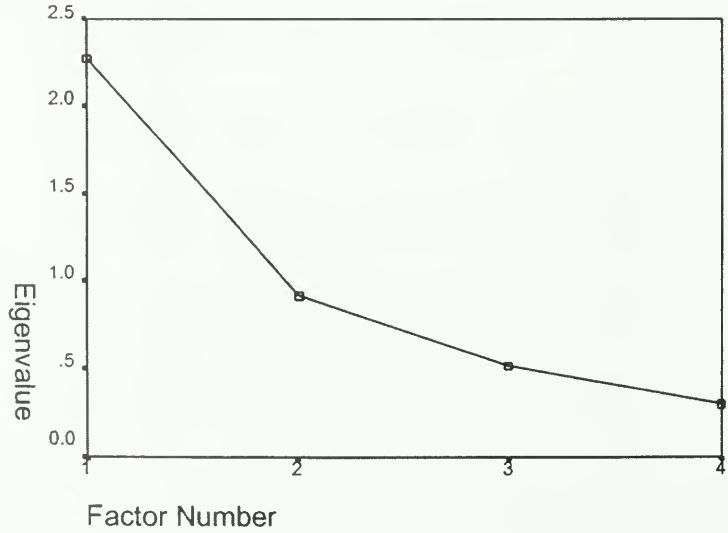
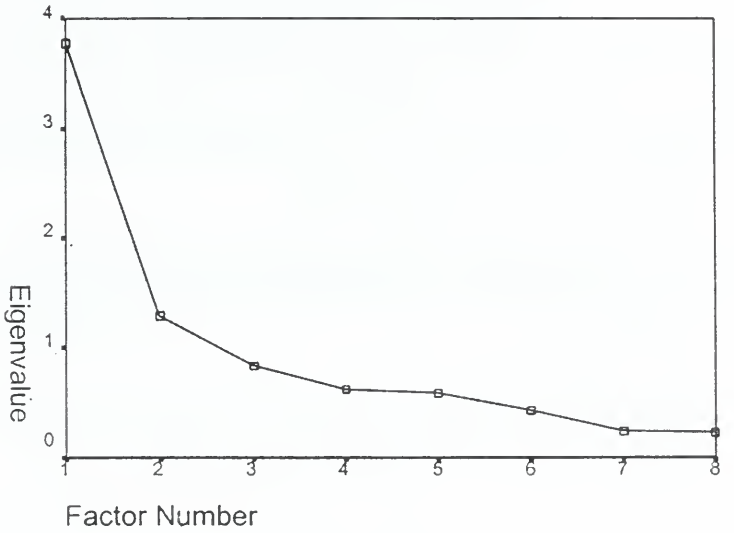


FIGURE A3 SCREE PLOT (MODEL 3)



UCLA ENGLISH AS A SECOND LANGUAGE PLACEMENT EXAM COMPOSITION RATING SCALE REVISED SEPTEMBER 1993

CONTENT

9-10 The essay fulfills the writing task well and treats the topic with sophistication. The main idea is clear and well-developed. Support is relevant, thorough and credible.

7-8 The essay addresses the writing task appropriately* and is developed competently. The main idea is clear and competently developed, but with less sophistication and depth than the 9-10 paper. Arguments/ideas are competently supported.

5-6 The essay addresses the writing task adequately, but may not be well-developed. OR The essay only addresses part of the topic, but develops that part sufficiently. The main idea is clear but may not be fully developed. Ideas/arguments may be unsupported or unrelated to main idea.

3-4 The essay only partially fulfills the writing task OR the main idea is somewhat clear, but requires the reader to work to find it. The essay contains unsupported or irrelevant statements.

1-2 The essay fails to fulfill the writing task and lacks a clear main idea and development. Most ideas/arguments are unsupported, and ideas are not developed. OR Not enough material to evaluate.

NOTE: *Appropriate* is defined as addressing all aspects of a writing topic, for example, all characteristics in questions involving choices. Furthermore, all parts of the writing task should be touched on in the writer's response.

RHETORICAL CONTROL

9-10 Introduction and conclusion effectively fulfill their separate purposes: The introduction effectively orients the reader to the topic and the conclusion not only reinforces the thesis but effectively closes off the essay.

Paragraphs are separate, yet cohesive, logical units which are well-connected to each other and to the essay\rquote s main idea. Sentences form a well-connected series of ideas.

7-8 The introduction presents the controlling idea, gives the reader the necessary background information, and orients the reader, although there may be some lack of originality in the presentation. The conclusion restates the controlling idea and provides a valid interpretation but not as effectively as the 9-10 paper.

Paragraphs are usually cohesive and logically connected to the essay's main idea. Sentences are usually well-connected.

5-6 Introduction presents the controlling ideas but may do so mechanically or may not orient the reader to the topic effectively. The conclusion does not give the reader new insights or may contain some extraneous information.
Paragraphs may exhibit a lack of cohesion or connection to the essay's main idea. Sentences may not be well-connected.

3-4 Introduction and conclusion do not restate the controlling idea. Introduction fails to orient the reader adequately, and the conclusion may be minimal or may not be tied to the rest of the essay. Paragraphs often lack cohesion and are not appropriately connected to each other or to the essay's main idea. Sentences are not well-connected.

1-2 Introduction and conclusion are missing or unrelated to rest of the essay.
There is no attempt to divide the essay into conceptual paragraphs, or the paragraphs are unrelated and the progression of ideas is very difficult to follow.
OR Not enough material to evaluate.

LANGUAGE (Grammar, Vocabulary, Register, Mechanics)

9-10 Except for rare minor errors (esp. articles), the grammar is native-like.
There is an effective balance of simple and complex sentence patterns with coordination and subordination. Excellent, near-native academic vocabulary and register. Few problems with word choices.

7-8 Minor errors in articles, verb agreement, word form, verb form (tense, aspect) and no incomplete sentences. Meaning is never obscured and there is a clear grasp of English sentence structure. There is usually a good balance of simple and complex sentences both appropriately constructed.

Generally, there is appropriate use of academic vocabulary and register with some errors in word choice OR writing is fluent and native-like but lacks appropriate academic register and sophisticated vocabulary.

5-6 Errors in article use and verb agreement and several errors in verb form and/or word form. May be some incomplete sentences. Errors almost never obscure meaning. Either too many simple sentences or complex ones that are too long to process. May be frequent problems with word choice; vocabulary is inaccurate or imprecise. Register lacks proper levels of sophistication.

3-4 Several errors in all areas of grammar which often interfere with communication, although there is knowledge of basic sentence structure.
No variation in sentence structure. Many unsuccessful subordinated or coordinated structures. Frequent errors in word choice (i.e. wrong word, not simply vague or informal. Register is inappropriate for academic writing.

1-2 There are problems not only with verb formation, articles, and incomplete sentences, but sentence construction is so poor that sentences are often incomprehensible. Sentences that are comprehensible are extremely simple constructions. Vocabulary too simple to express meaning and/or severe errors in word choice. OR Not enough material to evaluate.

ESLPE COMPOSITION RATING SCALE
REVISED FALL 1996

6 Exempt from ESL Service Courses

- Grammar is near native-like with little or no evidence of ESL errors. There may be basic writer developmental errors (e.g., spelling, sentence fragments & run-ons, interference from oral language)
- The writing exhibits a near native-like grasp of appropriate academic vocabulary and register and there are few problems with word choice OR the writing is fluent and native-like but lacks appropriate academic register or sophisticated vocabulary.
- Cohesion between paragraphs, sentences, and ideas is successfully achieved through a variety of methods (transitional words & phrases, a controlling theme, repetition of key words, etc.)

5 ESL 35

- The number and type of grammatical errors are limited and usually follow a discernible pattern; these include errors in article usage, noun number, subject/verb agreement, verb form (tense/aspect) and word form.
- Meaning is never obscured.
- The writing exhibits a variety of simple and complex sentence structures that are usually constructed appropriately, although there may be some problems with subordination or embedding.
- Register and vocabulary are generally appropriate to academic writing.
- Cohesion is adequate and achieved through the use of transitional words and phrases.

4 **ESL 33C**

- Grammar errors may occur in article usage and noun number, subject/verb agreement, verb form (tense/aspect), word form/choice, relative clause formation, passive voice, and coordination and subordination.
- Errors rarely obscure meaning.
- Sentence structure may range from too many simple sentences to complex ones that are too long to process. There may be some nonnative-like sentence fragments and run-ons.
- Vocabulary may be repetitive or inaccurate, and the register may exhibit a lack of academic sophistication.
- There may be a limited lack of cohesion and difficulty with paragraphing.

3 ESL 33B

- Patterns of errors occur in article usage and noun number, subject/verb agreement, verb form (tense/aspect), and/or word form.
- Errors occasionally obscure meaning.
- Although there is a good basic knowledge of sentence structure, there may be errors in or avoidance of relative clauses, passive voice, and/or coordination and subordination. There may be some nonnative-like sentence fragments and run-ons.
- Vocabulary may be repetitive and/or inaccurate. The register may be inappropriate at times.
- The writing exhibits a basic knowledge of cohesive devices but these may be misapplied, or the devices used may not create cohesion.

2 ESL 33A

- Frequent patterns of errors occur in article usage and noun number, subject/verb agreement, verb form (tense/aspect), and/or word form.
- Errors sometimes obscure meaning.
- Although there is a basic knowledge of sentence structure, there are errors in and avoidance/absence of relative clauses, passive voice, and/or coordination and subordination. There may be nonnative-like sentence fragments and run-ons.
- Vocabulary is generally basic and word choice is sometimes inaccurate. The register can often resemble either a conversational narrative or a stilted, confusing attempt at academic discourse.
- Although there is some use of cohesive devices, it is neither consistent nor always effective, and may be simple and repetitive in many cases.

1 ESL 832

- Pervasive patterns of errors occur in article usage and noun number, subject/verb agreement, verb form (tense/aspect), and word form.
- Except in very simple sentences, meaning is frequently obscured.
- A basic knowledge of sentence structure is lacking, and there are frequent errors in and/or avoidance of relative clauses, passive voice, and/or coordination and subordination. Nonnative-like sentence fragments and run-ons occur frequently.
- Vocabulary is quite basic, and more sophisticated attempts at word choice are often inaccurate or inappropriate. The register is often too conversational for academic purposes or, if an academic tone is attempted, it is incomprehensible.
- There may be attempts to use cohesive devices but they are either quite mechanical or so inaccurate that they mislead the reader.

0 No Response

ESLPE COMPOSITION RATING SCALE
REVISED FALL 2000

6 Exempt from ESL Service Courses

- Grammar is near native-like with little or no evidence of ESL errors. There may be basic writer developmental errors (e.g., spelling, sentence fragments & run-ons, interference from oral language)
- The writing exhibits a near native-like grasp of appropriate academic vocabulary and register and there are few problems with word choice OR the writing is fluent and native-like but lacks appropriate academic register or sophisticated vocabulary.
- Cohesion between paragraphs, sentences, and ideas is successfully achieved through a variety of methods (transitional words & phrases, a controlling theme, repetition of key words, etc.)

5 ESL 35

- The number and type of grammatical errors are limited and usually follow a discernible pattern; these include errors in article usage, noun number, subject/verb agreement, verb form (tense/aspect) and word form.
- Meaning is never obscured by grammar or lexical choices that are not native-like.
- The writing exhibits fluency, which is achieved through a variety of simple and complex sentence structures. These are usually constructed appropriately, although there may be some problems with more complex grammatical structures (e.g., subordination or embedding relative clauses).
- Register and vocabulary are generally appropriate to academic writing.
- Cohesion is adequate and achieved through the use of transitional words and phrases.

4 **ESL 33C**

- Grammar errors may occur in article usage and noun number, subject/verb agreement, verb form (tense/aspect), word form/choice, relative clause formation, passive voice, and coordination and subordination.
- Errors are noticeable, but rarely obscure meaning.
- The writing is less fluent than the 5 paper. It may sound choppy because there are too many simple sentences. Or there may be too many complex sentences that are too long to process. Or there may be some non-native-like sentence fragments and run-ons.
- Vocabulary may be repetitive or inaccurate, and the register may exhibit a lack of academic sophistication.
- There may be a limited lack of cohesion and difficulty with paragraphing.

3 ESL 33B

- Patterns of errors occur in article usage and noun number, subject/verb agreement, verb form (tense/aspect), and/or word form.

- Errors are noticeable and occasionally obscure meaning.
- Although there is a good basic knowledge of sentence structure, the writing lacks fluency because of errors in or avoidance of relative clauses, passive voice, and/or coordination and subordination. There may be some non-native-like sentence fragments and run-ons.
- Vocabulary is inaccurate in places and may rely on repeating words and expressions from the prompt.
- The writing exhibits a basic knowledge of cohesive devices but these may be misapplied, or the devices used may not create cohesion.

2 **ESL 33A**

- Frequent patterns of errors occur in article usage and noun number, subject/verb agreement, verb form (tense/aspect), and/or word form.
- Errors are noticeable and obscure meaning.
- Although there is a basic knowledge of sentence structure, this is not consistently applied. The writing exhibits errors in and avoidance/absence of relative clauses, passive voice, and/or coordination and subordination. There are non-native-like sentence fragments and run-ons.
- Vocabulary is generally basic and word choice is inaccurate. The writer may rely on repeating words or expressions from the prompt. The register can often resemble either a conversational narrative or a stilted, confusing attempt at academic discourse.
- Although there is some use of cohesive devices, it is neither consistent nor always effective, and may be simple and repetitive in many cases.

1 Pre-University

- Pervasive patterns of errors occur in article usage and noun number, subject/verb agreement, verb form (tense/aspect), and word form.
- Except in very simple sentences, meaning is frequently obscured.
- A basic knowledge of sentence structure is lacking, and there are frequent errors in and/or avoidance of relative clauses, passive voice, and/or coordination and subordination. When sentences are complete, they are often simple or are expressions learned as "chunks".
- Vocabulary is quite basic, and more sophisticated attempts at word choice are often inaccurate or inappropriate. The register is often too conversational for academic purposes or, if an academic tone is attempted, it is incomprehensible.
- There may be attempts to use cohesive devices but they are either quite mechanical or so inaccurate that they mislead the reader.

0 No Response

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, 34, 21-42.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-31). Urbana, IL: NCTE.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7, 31-51.

Cumming, A., & Mellow, D. (1996). An investigation into the validity of written indicators of second language proficiency. In A. Cumming & R. Berwick (Eds.), *Validation in language testing. Modern languages in practice 2* (pp. 72-93). Bristol, PA: Multilingual Matters.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher*, 27 (2), 14-24.

Fathman, A. K., & Whalley, E. (1990). Teacher response to student writing: Focus on form versus content. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 178-190). Cambridge: University of Cambridge Press.

Gay, L. R. (1992). *Educational research: Competencies for analysis and application*. New York: Macmillan.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge: University of Cambridge Press.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-278). Norwood, NJ: Multilingual Matters.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly*, 29, 759-762.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, 41, 337-373.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000—Writing: Composition, community, and assessment. TOEFL Monograph Series MS-5*. Princeton: Educational Testing Service.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly*, 18, 87-107.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, 41, 201-213.

Isaac, S., & Michael, W. B. (1995). *Handbook in research and evaluation (3rd ed.)*. San Diego, CA: Educational and Industrial Testing Services.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Johns, A. M. (1986). Coherence and academic writing: Some definitions and suggestions for teaching. *TESOL Quarterly*, 20, 247-265.

Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 140-154). Cambridge: University of Cambridge Press.

Lewis-Beck, M. S. (1980). *Applied Regression: An introduction* (Sage University Paper Series on Quantitative Applications in the Social Sciences No. 07-022). Newbury Park, CA: Sage.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching*. Englewood Cliffs, NJ: Merrill.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment. Studies in Language Testing 3* (pp. 92-114). Cambridge: University of Cambridge Press.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models (4th ed.)*. Chicago: Irwin.

Popham, W. J. (1997). What's wrong—and what's right—with rubrics. Educational Leadership, 55(2).

SPSS Inc. (1997). *SPSS for Windows Release 8.0.0 Standard Version*. Chicago: SPSS Inc.

Tyndall, B., & Kenyon, D. M. (1996). Validation of a new holistic rating scale using Rasch multi-faceted analysis. In A. Cumming & R. Berwick (Eds.), *Validation in language testing. Modern Languages in Practice 2* (pp. 39-57). Bristol, PA: Multilingual Matters.

Vacc, N. N. (1989). Writing evaluation: Examining four teachers' holistic and analytic scores. *The Elementary School Journal*, 90(1), 87-95.

Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-126). Norwood, NJ: Ablex.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11, 197-223.

Weigle, S. C., & Lynch, B. (1996). Hypothesis testing in construct validation. In A. Cumming & R. Berwick (Eds.), *Validation in language testing. Modern languages in practice 2* (pp. 58-71). Bristol, PA: Multilingual Matters.

White, E. M. (1984). Holisticism. *College Composition and Communication*, 35, 400-409.